Recommendation of Web Pages using Weighted K-Means Clustering

R. Thiyagarajan
Department of Computer Applications
Nehru Institute of IT and Management
Coimbatore-641 105. India

K. Thangavel
Department of Computer Science
Periyar University
Salem-636 011. India

R. Rathipriya
Department of Computer Science
Periyar University
Salem-636 011. India

ABSTRACT
Web Recommendation Systems are implemented by using collaborative filtering approach. It is a specific type of information filtering system that aims to predict the user browsing activity and then recommend to the user web pages items that are likely to be of interest. In this paper, a new recommendation system is proposed by using Weighted K-Means clustering approach to predict the user's navigational behavior. The proposed recommendation system based on Weighted K-Means clustering performs well when compared to K-Means algorithm. The performance of the comparative analysis is presented through experimental results.

Keywords
Web Usage Mining, Web recommendation system, K-Means clustering, Weighted K-Means clustering, Hamming distance, Mean square residue.

1. INTRODUCTION
Nowadays Web becomes the backbone of information. The major problem for the internet users is being unable to retrieve useful and relevant information. The browsing patterns of the users can help organizations to recommend the more relevant web pages according to the current interests of the user.

Web Usage Mining (WUM) mines user access patterns from usage logs, which record clicks made by every user. The output of WUM is some patterns that may be the input to the recommendation systems which is one of the application areas of the web usage gives the ability to predict the next visited page for a given user [6]. The main goal of the recommendation system is to improve the web site usability by knowing the interest of the users. The web recommendation process consists of two components namely online and off-line with respect to web server activity. Offline component builds the knowledge base by analyzing historical data, such as server access log file or web logs which are captured from the server. Then these web logs are used in the online component for capturing the intuition list of the user so as to recommend page views to the user whenever user comes online for the next time [2].

In this paper, a framework is generated for capturing recommendations in the form of recommendation list for user using Weighted K-Means clustering. A recommendation list consists of list of pages visited by user as well as list of pages visited by other users of having similar usage profile. The rest of this paper is organized as follows: In section 2, recommendation system using web usage mining is discussed. Section 3 presents the block diagram and the implementation for the usage based recommendation system using K-Means and Weighted K-Means Clustering algorithms. Results and discussion are revealed in section 4. Finally, section 5 concludes the paper with the direction for future work.

2. RELATED WORK
At present many researchers have proposed various recommendation systems for online personalization through web usage mining [2]. This kind of recommendation system is used to predict the user navigation behavior and their preferences using web log data.

Bamshad Mobasher [1] presented a system called Web personalizer which provides dynamic recommendations, as a list of hypertext links, to users. In preprocessing phase, the data mining techniques (i.e. clustering, sequence pattern discovery and association rules) are used to obtain the aggregate usage profiles. In offline phase web server logs are converted into clusters of visited pages, and cluster made up of set of pages with common usage characteristics or behavior. In the online phase, active user session is considered in order to find matches among users' activities and discovered similar usage profiles. Matching usage profiles are used to compute a set of recommendations which will be inserted into last requested page as list of hypertext links. Sumathi et al. [2] introduced the recommender systems based on the user's navigational patterns using model based clustering and suitable recommendations has been provided to cater to the needs of the user. AlMurtadha et al. [3] have focused on improving the prediction of the next visited web pages and recommends it to the current anonymous user by assigning them to the best navigation profiles obtained by previous navigations of similar interested users. NEWER, a usage-based Web recommendation system presented by Castellano, Fanelli and Torsello [4] exploits the potential of Computational Intelligence techniques to suggest dynamically interesting pages to users according to their preferences. Almurtadha, Sulaiman, Mustapha and Udzir [5] focused on IPACT, an improved recommendation system using Profile Aggregation based on Clustering of Transactions. In [14], Fuzhi ZHANG, Huilin LIU and jinbo CHAO presented A Two-stage Recommendation Algorithm based on K-means Clustering in Mobile E-commerce. K.Thangadurai, M.Uma and M.Punithavalli [12] had a Study on Rough Clustering in which rough K-means clustering is studied and compared with the traditional K-means and weighted K-Means clustering methods for different data sets available in UCI data repository

3. USAGE BASED RECOMMENDATION SYSTEM
The proposed framework consists of two main components, namely the offline and online as shown in figure 1.

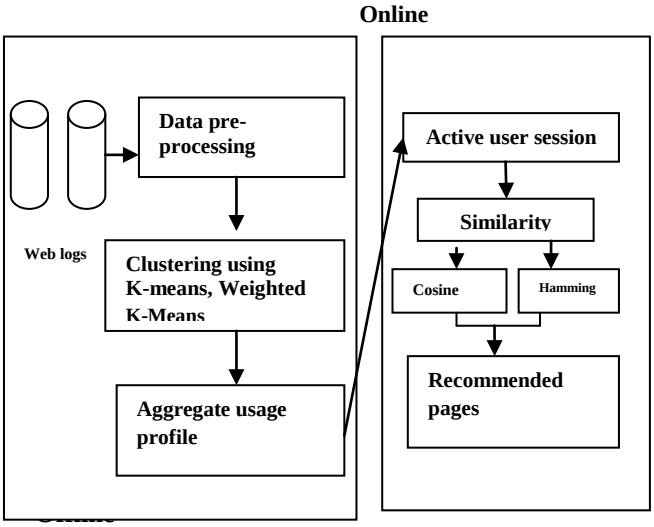

Figure 1: Framework of Recommendation System for online users

In the offline component the three important steps are considered. First step is to preprocess the web server logs or web usage data by applying data cleaning techniques and then partition the web navigations into sessions determined by the period of browsing. Second one is to partition the filtered sessionized page views into clusters of users navigation patterns with similar page views browsing activities using K-Means algorithm[7] and Weighted K-Means algorithm [12]. Finally, web navigation profiles are generated based on the preformed clusters. The online component does the matching of the new anonymous user request (current active session) to the profile shares common interests to the user.

The usage profile contains only those web pages that passed certain confidence support and weights values. The confidence support determines the frequent occurrence on those pages in the cluster. These profiles don't consider specific users, since this study don't consider the users history in account during the profile generation. The usage profile is constructed as a set of pageview and its weight as pair using equation (1).

$$\textit{Usage profile} = \{(p, \textit{weight}(p)) \mid p \in P, \textit{weight}(p) \geq \textit{min\_weight} \}. \quad (1)$$

where $P = \{p_1, p_2, \ldots, p_n\}$, a set of *n* pageviews appearing in the transaction file with each pageview uniquely represented by its associated Uniform Resource Locator (URL) and the weight(p) is the (mean) value of the attribute's weights in the cluster.

3.1 Preprocessing of Click stream Data

Click stream data means that when a user viewed a sequence of web pages then web pages are displayed one by one on a row at a time [10, 11]. Analysis of clicks is the process of extracting knowledge from web logs. This analysis involves data preprocessing and then applying data mining techniques. Data preprocessing involves data extraction, cleaning and filtration followed by identification of their sessions.

3.2 K-Means Clustering Algorithm

Even though there are quite number of algorithms for clustering, the bench mark K-Means clustering technique [7] has been used to group the web users in this paper. Consider a data set contain the data to be clustered data point, D = {X_1… Xn}, first choose from this data points, K initial centroids randomly, where K is user-parameter, the number of clusters desired. It uses an iterative hill-climbing algorithm. The process of K-means clustering is explained as follows [13]:

(i) The initial seeds with the chosen number of clusters, K, are selected and an initial partition is built by using the seeds as the centroids of the initial clusters.
(ii) Each data point is assigned to the centroid that is nearest, thus forming a cluster.
(iii) Keeping the same number of clusters, the new centroid of each cluster is calculated.
(iv) Iterate Steps (ii) and (iii) until the clusters stop changing or stop conditions are satisfied.

In K-Means Clustering, all the pages are equally considered, but some of the pages have been visited by more number of users. To consider the pages which are visited by more number of users, the suitable weight is assigned to each page in order to give more importance for that page at the time of clustering. Hence, Weighted K-Means algorithm has been proposed in this paper and web page recommendation have been done accordingly. In this paper, Weighted K-Means clustering is applied in the offline phase to generate the similar user groups based on their usage behavior, since these user groups or user clusters are used to generate the usage profile using equation (1).

3.3 Weighted K-Means algorithm

Weighted K-Means algorithm [12] is one of the clustering algorithms, based on the K-Means algorithm calculating with weights. A natural extension of the K-Means problem allows us to include some more information, namely, a set of weights associated with the data points. These might represent a measure of importance, a frequency count, or some other information. Weighted K-Means attempts to decompose a set of objects into a set of disjoint clusters, taking into consideration the fact that the numerical attributes of objects in the set often do not come from independent identical normal distribution.

The weighted K-Means algorithm uses weight vector to decrease the affects of irrelevant attributes and reflect the semantic information of objects. Weighted K-Means algorithm is iterative and use hill-climbing to find an optimal solution (clustering), and thus usually converge to a local minimum.
In this algorithm, the weights can be classified into two types. ***Dynamic Weights:*** In the dynamic weights, the weights are changed during the program. ***Static Weights:*** In the static weights, the weights are not changed during the program. The Weighted K-Means algorithm is used to cluster the objects.

The working procedure of **Weighted K-Means clustering** is as follows.

Input: A set of n data points and the number of clusters (K)

Output: Centroids of the K clusters

(i) Initialize the number of clusters k.
(ii) Randomly selecting the centroids ($c_1, c_2, \ldots, c_K$) in the data set.
(iii) Choosing the Static weight w ,which is range from 0 to 2.5 or (5.0)
(iv) Find the distance between the centroids using the Euclidean Distance equation. $d_{ij} = w.*(x_i - c_k)^2$

(v) Update the centroids using this equation.
(vi) Stop the process when the new centroids are nearer to old one. Otherwise, go to step-(iv).

3.4 Cosine Similarity [2,8]

The similarity of the active session with each of the discovered aggregate profile is determined using the well-known Cosine similarity measure. If an active session s_i is taken from cluster c_k, then their similarity can be measured as follows:

$$\mathrm{sim}(s_j, c_k) = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \sqrt{\sum_{i=1}^{n} w_{i,k}^2}}$$

where $w_{i,j}$ represents weight of page i in active session j and $w_{i,k}$ represents weight of page i in cluster k.

3.5 Hamming Similarity [9]

Given a space of vectors, the Hamming distance between two vectors is defined as the number of components in which they differ. It should be obvious that Hamming distance is a distance measure. Clearly the Hamming distance cannot be negative, and if it is zero, then the vectors are identical. Most commonly, Hamming distance is used when the vectors are binary; they consist of 0's and 1's only. However, in principle, the vectors can have components from any set. For example the Hamming distance between the vectors 10101 and 11110 is 3. That is, these vectors differ in the second, fourth, and fifth components, while they agree in the first and third components.

Hamming Similarity (u_i, u_j) = 1- Hamming distance(u_i, u_j)

In the first phase usage profiles are extracted. In second phase, two different similarity measures namely Cosine similarity and Hamming similarity measures are used to measure the similarity between the active user and the extracted usage profiles. The recommendation list is generated from the nearest usage profiles. Usage profiles whose similarity value greater than the threshold μ (here μ=0.5) are considered as the nearest profile to the given active user.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Data Set

A real dataset is used for this experiment. The data set is taken from the UCI dataset repository (http://kdd.ics.uci.edu/) that consists of Internet Information Server (IIS) logs for *msnbc.com* and news-related portions of *msn.com* for the entire day of September 28, 1999 (Pacific Standard Time). Visits are recorded at the level of URL category and are recorded in time order. Each sequence in the dataset corresponds to page views of a user during that twenty-four hour period. Each event in the sequence corresponds to a user's request for a page. Requests are not recorded at the finest level of detail that is, at the level of URL, but rather, they are recorded at the level of page category (as determined by a site administrator). The categories are "front page", "news, "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "msn-news", and "msn-sports". Any page requests served via a caching mechanism were not recorded in the server logs and hence, not present in the data. This dataset is slightly changed flattering to our experiment, if the user visit only the "front page" then 1 is recorded on the first position of the matrix and other 16 column (category) are filled by 0 [2]. The details of the data set are provided in Table 1.

Table 1. Dataset used in the experiment

Dataset	***MSNBC***
Total Number of Users	989818
Average number of visit per user	5.7
Number of URL for each categories	10-5000

Users visited more than 8 page view categories are considered for recommendation. Therefore, the number of user sessions after this data filter step is 3408. In the first phase, K-Means clustering technique and Weighted K-Means clustering techniques are applied to the MSNBC dataset with K=10.

Table 2. List of MSR values and page view categories using Weighted K-Means and K-Means of 10 clusters

Cluster Index	**Weighted K-Means MSR**	**Pageview Categories**	**Normal K-Means MSR**	**Pageview Categories**
1	46.8082	1 2 3 4 5 6 7 10 11 12 14	39.0137	1 2 3 4 6 7 10 12 15
2	45.4358	1 2 3 4 5 6 7 10 11 12 14	34.0003	1 2 3 4 6 10 11 12 14 17
3	50.6072	1 2 3 4 5 6 7 10 11 12 14	39.2458	1 2 3 4 5 6 7 10 11 15
4	44.4727	1 2 3 4 5 6 7 10 11 12 14 15	42.9592	1 2 4 6 7 8 9 14
5	47.7274	1 2 3 4 5 6 7 10 11 12 14	32.1258	1 2 3 4 5 6 7 10 11
6	42.71	1 2 3 4 5 6 7 10 11 12 14 15	35.2479	1 2 4 6 7 10 11 12 14
7	45.3843	1 2 3 4 5 6 7 10 11 12 14	40.8253	1 2 4 6 7 9 11 12
8	43.6248	1 2 3 4 5 6 7 10 11 12	47.0673	1 2 3 4 6 7 9 12 13 14

		14		
9	42.1685	1 2 3 4 6 7 10 11 12 14 17	15.5187	1 2 3 4 5 6 7 8 10 11 12 14 15
10	42.5192	1 2 3 4 5 6 7 8 10 11 12 14 15	36.2595	1 2 3 4 5 6 10 11 12 14

Table 2 shows the Mean Square Residue value of the 10 clusters and pageview categories using K-Means and Weighted K-Means in the usage profile for the corresponding clusters. In the second phase, active user visits the 1 and 2 pageview categories. It is symbolically denoted as
A=[1, 1, 0 , 0, 0 , 0 , 0, 0, 0 ,0 , 0, 0 , 0 , 0, 0 , 0 , 0, 0]

Table 3 tabulates the Cosine similarity and Hamming similarity value for the 10 clusters using K-Means clustering with respect to the active user sessions.

Table 3. Similarity measures using K-Means clustering

Cluster Index	Cosine Similarity	Hamming Similarity
1	1	0.8235
2	0.6493	0.8824
3	0.974	0.8235
4	0.6866	0.8824
5	1	0.8235
6	0.6667	0.8824
7	0.6866	0.8824
8	0.974	0.8235
9	0.9122	0.8235
10	0.6493	0.8824
Average Similarity	0.81987	0.85295

Visited Pages = (3 6 7 9 10 12 13 14) which is denoted as A=[0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0]

Figure 2 shows the graphical representation of the comparison of Cosine and Hamming similarity measures for the clusters generated using K-Means and Weighted K-Means clustering techniques. It is clearly depicts that Hamming similarity value is higher than the cosine similarity value. This is because of the binary representation of the web data. Table 4 exhibits Cosine similarity and Hamming similarity value for the 10 clusters using Weighted K-Means clustering with respect to the active user sessions is given below.

Table 4. Similarity measures using Weighted K-Means clustering

Cluster Index	Cosine Similarity	Hamming Similarity
1	0.9511	0.9991
2	0.9511	0.9991
3	0.9306	0.9991
4	0.974	0.9991
5	0.9511	0.9991
6	0.9511	0.9991
7	0.9511	0.9991
8	0.8954	0.9991
9	0.9122	0.9991
10	0.9511	0.9991

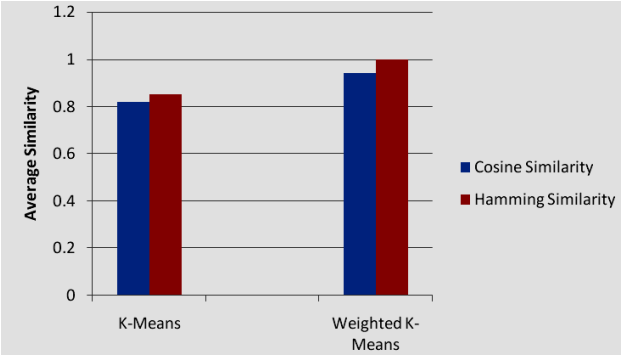

Figure 2: Comparison of Similarity Measures

Table 5. Recommendations using K-Means clustering

Similarity Measure	List of Recommended Pages	Recommendation Quality Percentage
Cosine	1 2 4 5 8 9 10 11 12 13 14 15	83.33
Hamming	1 2 4 5 8 9 10 11 12 13 14 15	83.33

Table 6. Recommendations using Weighted K-Means clustering

Similarity Measure	List of Recommended Pages	Recommendation Quality Percentage
Cosine	1 2 4 5 7 8 9 10 11 14 15 16	66.67
Hamming	1 2 4 7 8 9 10 11 12 13 14 16	100%

Percentage of Recommendation Quality =Number of correctly recommended pages/ (Total Number of Visited pages- Number of Pages in the Active User Session) *100

Table 5 and Table 6 show the recommendation list for the active user as given above using traditional K-Means clustering and Weighted K-Means clustering respectively. The distance measure used in the both clustering methods is Euclidean distance. From this empirical study, it has observed that Hamming similarity using Weighted K-Means clustering gives better recommendation quality than cosine similarity measure for the binary web usage data. The list of recommended pages is provided in Table 6.

5. CONCLUSION

This paper has paid an attention to group the similar usage behavior of users using Weighted K-Means algorithm for aggregated usage profile and new validating measure called MSR(Mean Square Residue) is applied to evaluate the cluster's quality. The results of this clustering approach are compared with the results of traditional clustering called K-Means. It was observed that the usage profile extracted from the MSNBC dataset using Weighted K-Means provides high quality recommendation for the given active user than results obtained by using K-Means Clustering. In future, the overlapping clusters may be obtained and these clusters may be used for usage profile generation. Hence more pages visited by users can be considered for recommendation process.

6. REFERENCES

[1] Bamshad Mobasher,2001. WebPersonalizer: A Server-Side Recommender System Based on Web Usage Mining. Technical ReportTR01-010,Schoolof computer Science, telecommunications and Information Systems, DePaul University, Chicago, IL, USA

[2] C.P. Sumathi et. al. / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 09, 2010, 3046-3052

[3] AlMurtadha, Y.M., M.N.B. Sulaiman, N. Mustapha and N.I. Udzir,2010. Mining web navigation profiles for recommendation system. Inform.Technol. J., 9: 790-796. DOI:10.3923/itj.2010.790.796

[4] Castellano, G., Fanelli, A.M., & Torsello, M.A. (2011). NEWER: A system for NEuro-fuzzy Web recommendation. Applied soft Computing,11(1), 793-806.

[5] AlMurtadha, Y.,Sulaiman, M..N.B., N. Mustapha and N.I. Udzir,(2011). IPACT: Improved web page recommendation System Using Profile Aggregation Based on Clustering of Transactions, American Journal of Applied Sciences, 8(3),277-283.

[6] Yahya AlMurtadha, Md. Nasir Sulaiman, Norwati Mustapha and Nur Izura Udzir. Improved web page recommender System Based on Web Usage Mining, Proceedings of the 3rd International Conference on Computing and Informatics, ICOCI 2011, 8-9 June 2011 Bandung, Indonesia, Paper No. 079.

[7] Ms. Vinita Shrivastava, Mr. Neetesh GuptaPerformance Improvement Of Web Usage Mining By Using Learning Based K-Mean Clustering on International Journal of Computer Science and its Applications-[ISSN 2250 - 3765].

[8] Khribi, M. K., Jemni, M., & Nasraoui, O. Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval (2009) , Educational Technology & Society, 12 (4), 30–42.

[9] F. Khalil, J. Li, H. Wang. An Integrated Model for Next Page Access Prediction, Copyright °c 2009 Inderscience Enterprises Ltd.

[10] Cooley, R., B. Mobasher and J.Srivatsava, 1997. Web mining information and pattern discovery on the world wide web. Proceeding of the 9th IEEE International Conference on tools with Artificial Intelligence, Newport Beach,CA., pp: 558-567. DOI: 10.1109/TAI.1997.632303.

[11] Ms.Dipa Dixit, Mr.Jayant Gadge, Automatic Recommendation for Online Users Using Web Usage Mining on International Journal of Managing Information Technology (IJMIT) Vol.2, No.3, August 2010

[12] Dr.K.Thangadurai, M.Uma, Dr.M.Punithavalli, A Study On Rough Clustering Global Journal of Computer Science and Technology Vol. 10 Issue 5 Ver. 1.0 July 2010 P a g e | **55**

[13] Kyoung-jae Kim, Hyunchul Ahn, A Recommender system using GA K-means clustering in an online shopping market.., Expert Systems with Applications (20070, doi:10,1016/j.eswa.2006.12.025.

[14] Fuzhi ZHANG, Huilin LIU, jinbo CHAO, A Two-stage Recommendation Algorithm based on K-means Clustering in Mobile E-commerce, Journal of Computational Information Systems 6:10 (2010) 3327-3334.